EVALUASI PERFORMA *SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE* (SMOTE) UNTUK MENGATASI KLASIFIKASI DATA TIDAK SEIMBANG PADA METODE *K-NEAREST NEIGHBOR* (KNN) DAN *SUPPORT VECTOR MACHINE* (SVM)

(Skripsi)

Oleh

WIDYA AMALIA PUTRI RISWANDHA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2023

ABSTRACT

EVALUATION OF SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE (SMOTE) PERFORMANCE TO OVERCOME UNBALANCED DATA CLASSIFICATION IN K-NEAREST NEIGHBOR (KNN) AND SUPPORT VECTOR MACHINE (SVM) METHODS

OLEH

WIDYA AMALIA PUTRI RISWANDHA

K-Nearest Neighbor (KNN) and Support Vector Machine (SVM) methods are one of the most popular techniques in classification. In general, a common problem with classification is the presence of data imbalances. Data imbalance is a condition in which the distribution in the number of classes is uneven or can be interpreted as having a significant difference in the number of classes. It affects the performance of the classification results. So from that, it's important to address data imbalances, one of which is using the Synthetic Minority Oversampling Technique. (SMOTE). The study aims to study and evaluate the performance of SMOTE to address the imbalanced classification of data in KNN and SVM. Based on the results of the analysis obtained that SMOTE is effective in improving the performance of the classification of diabetic patients which is proven by the presence of increased accuracy value in the method KNN with parameter $k = 9$ obtaining the accurate value of 81.58% and in the SVM method with the RBF kernel obtaining the accurate value of 85,53%.

Keywords: Classification, Imbalance Data, KNN, SVM, SMOTE

ABSTRAK

EVALUASI PERFORMA *SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE* (SMOTE) UNTUK MENGATASI KLASIFIKASI DATA TIDAK SEIMBANG PADA METODE *K-NEAREST NEIGHBOR* (KNN) DAN *SUPPORT VECTOR MACHINE* (SVM)

OLEH

WIDYA AMALIA PUTRI RISWANDHA

Metode *K-Nearest Neighbor* (KNN) dan *Support Vector Machine* (SVM) merupakan salah satu teknik yang populer pada klasifikasi. Pada umumnya, masalah yang sering terjadi pada klasifikasi yaitu adanya ketidakseimbangan data. Ketidakseimbangan data merupakan suatu keadaan dimana distribusi pada jumlah kelas tidak merata atau dapat diartikan bahwa terdapat perbedaan yang signifikan terhadap jumlah kelas. Hal ini mempengaruhi terhadap performa hasil klasifikasi. Maka dari itu, penting untuk mengatasi masalah ketidakseimbangan data, salah satunya dengan menggunakan *Synthetic Minority Oversampling Technique* (SMOTE). Penelitian ini bertujuan untuk mengkaji dan mengevaluasi performa SMOTE untuk mengatasi klasifikasi data tidak seimbang pada KNN dan SVM. Berdasarkan hasil analisis diperoleh bahwa SMOTE efektif untuk meningkatkan performa klasifikasi penderita diabetes yang dibuktikan dengan adanya peningkatan nilai akurasi pada metode KNN dengan parameter $k = 9$ didapatkan nilai akurasi sebesar 81.58% dan pada metode SVM dengan Kernel RBF didapatkan nilai akurasi sebesar 85.53%.

Kata Kunci: Klasifikasi, Ketidakseimbangan Data, KNN, SVM, SMOTE

EVALUASI PERFORMA *SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE* (SMOTE) UNTUK MENGATASI KLASIFIKASI DATA TIDAK SEIMBANG PADA METODE *K-NEAREST NEIGHBOR* (KNN) DAN *SUPPORT VECTOR MACHINE* (SVM)

Oleh

WIDYA AMALIA PUTRI RISWANDHA
1917031036

Skripsi

Sebagai Salah Satu Syarat untuk Memperoleh Gelar
SARJANA MATEMATIKA

Pada

Jurusan Matematika
Fakultas Matematika dan Ilmu Pengetahuan Alam

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2023

Judul Skripsi	: **EVALUASI PERFORMA *SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE* (SMOTE) UNTUK MENGATASI KLASIFIKASI DATA TIDAK SEIMBANG PADA METODE *K-NEAREST NEIGHBOR* (KNN) DAN *SUPPORT VECTOR MACHINE* (SVM)**
Nama Mahasiswa	: Widya Amalia Putri Riswandha
Nomor Pokok Mahasiswa	: **1917031036**
Jurusan	: **Matematika**
Fakultas	: **Matematika dan Ilmu Pengetahuan Alam**

MENYETUJUI

1. Komisi Pembimbing

Dr. Khoirin Nisa, S.Si., M.Si
NIP. 197407262000032001

Dra. Dorrah Aziz, M.Si.
NIP. 196101281988112001

2. Ketua Jurusan Matematika

Dr. Aang Nuryaman, S.Si., M.Si.
NIP. 197403162005011001

MENGESAHKAN

1. Tim Penguji

   Ketua : **Dr. Khoirin Nisa, S.Si., M.Si.** .....

   Sekretaris : **Dra. Dorrah Aziz, M.Si.** .....

   Penguji : **Prof. Ir. Netti Herawati, M.Sc., Ph.D.** .....
   Bukan Pembimbing

2. Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam

Dr. Eng. Heri Satria, S.Si., M.Si.
NIP. 197110012005011002

Tanggal Lulus Ujian Skripsi : **14 Agustus 2023**

PERNYATAAN SKRIPSI MAHASISWA

Yang bertanda tangan di bawah ini:

Nama	: **Widya Amalia Putri Riswandha**
Nomor Pokok Mahasiswa	: **1917031036**
Jurusan	: **Matematika**
Judul	: **Evaluasi Performa *Synthetic Minority Oversampling Technique* (SMOTE) untuk Mengatasi Klasifikasi Data Tidak Seimbang pada Metode *K-Nearest Neighbor* (KNN) dan *Support Vector Machine* (SVM)**

Dengan ini menyatakan bahwa penelitian ini adalah hasil pekerjaan saya sendiri dan apabila di kemudian hari terbukti bahwa skripsi ini merupakan hasil salinan atau dibuat oleh orang lain, maka saya bersedia menerima sanksi sesuai dengan ketentuan akademik yang berlaku.

Bandar Lampung, 14 Agustus 2023

Penulis,

Widya Amalia Putri Riswandha

RIWAYAT HIDUP

Penulis bernama lengkap Widya Amalia Putri Riswandha. Lahir di Pandeglang pada 15 Januari 2002. Penulis merupakan anak pertama dari empat bersaudara, dari pasangan Bapak Wahyu Supardan dan Ibu Ratu Entin Supartini.

Penulis mengawali pendidikan taman kanak-kanak di TK Bhayangkara Pandeglang yang diselesaikan pada tahun 2007. Kemudian penulis melanjutkan pendidikan sekolah dasar di SDN 3 Pandeglang yang diselesaikan pada tahun 2013. Selanjutnya penulis melanjutkan pendidikan sekolah menengah pertama di MTsN 1 Pandeglang yang diselesaikan pada tahun 2016 dan melanjutkan pendidikan sekolah menengah atas di SMAN 1 Pandeglang yang diselesaikan pada tahun 2019.

Pada tahun 2019 penulis melanjutkan pendidikan Strata Satu (S1) di Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung melalui jalur SNMPTN. Selama berstatus mahasiswa, penulis berkesempatan untuk aktif di beberapa organisasi yaitu Anggota Bidang Kaderisasi dan Kepemimpinan Himpunan Mahasiswa Matematika (HIMATIKA) FMIPA Unila periode 2020, Staf Ahli Kementrian Advokasi Publik BEM U KBM Universitas Lampung periode 2020, Bendahara Dinas Hubungan Luar BEM FMIPA Universitas Lampung periode 2021, dan Sekretaris Dinas Hubungan Internal dan Eksternal BEM FMIPA Universitas Lampung periode 2022.

Pada bulan Januari hingga Februari 2022 penulis melaksanakan Kerja Praktik (KP) di Badan Pendapatan Daerah Provinsi Lampung sebagai bentuk

pengembangan diri serta menerapkan ilmu yang telah didapat selama perkuliahan. Pada bulan Juni hingga Agustus 2022 penulis melaksanakan Kuliah Kerja Nyata (KKN) Periode II di Desa Wana, Kecamatan Melinting, Kabupaten Lampung Timur sebagai bentuk pengabdian kepada masyarakat. Selama menjadi mahasiswa, penulis juga berkesempatan untuk mengikuti program MBKM yaitu Kampus Mengajar di SDN 2 Kabayan Pandeglang pada bulan Agustus hingga Desember 2021.

KATA INSPIRASI

"Allah tidak membebani seseorang melainkan sesuai dengan kesanggupannya..."
(Q.S Al-Baqarah: 286)

"...dan aku belum pernah kecewa dalam berdoa kepada-Mu, ya Tuhanku"
(Q.S Maryam: 4)

"Selalu ada harga dalam sebuah proses. Nikmati saja lelah-lelah itu. Lebarkan lagi rasa sabar itu. Semua yang kau investasikan untuk menjadi dirimu serupa yang kau impikan. Mungkin tidak akan selalu lancar, tapi gelombang-gelombang itu yang nanti bisa kau ceritakan."
(Boy Chandra)

"Tidak apa-apa jika kamu berjalan perlahan, asalkan tidak berhenti di tengah jalan. Jika lelah, beristirahatlah sejenak. Tidak apa-apa jika saat ini jalanmu terasa sangat sulit dan setiap langkah yang kamu ambil terasa sangat berat. Berjalanlah selangah demi selangkah, meski dengan mata sembab sebab menangisi rasa takut, khawatir, cemas, dan pikiran jahat setiap malah. Selama kamu masih melangkah maju dan tidak menyerah, kamu harus yakin bahwa Tuhan akan selalu menyertai kita dalam segala hal. "
(Widya Amalia Putri Riswandha)

PERSEMBAHAN

Dengan mengucapkan puji dan syukur saya haturkan kepada Allah SWT. yang telah memberikan rahmat, hidayah, dan karunia-Nya kepada saya. Kupersembahkan karya sederhana dengan penuh ketulusan hati kepada orang yang ku kasihi dan ku sayangi sebagai wujud rasa cinta dan sayangku kepada:

Bapak Wahyu Supardan dan Ibu Ratu Entin Supartini Tercinta

Sebagai tanda bakti, hormat, dan rasa terima kasih yang tiada terhingga kupersembahkan karya kecil dan sederhana ini kepada Papah (Wahyu Supardan) dan Mamah (Ratu Entin Supartini) yang telah memberikan kasih sayang, dukungan, ridho, dan cinta kasih yang tidak terhingga yang tidak mungkin dapat kubalaskan hanya dengan selembar kertas yang bertuliskan kata persembahan. Semoga ini menjadi langkah awal untuk membuat Papah dan Mamah bahagia dan bangga terhadapku. Terima kasih sudah menjadi orang tua yang sangat amat hebat untukku.

Dosen Pembimbing dan Pembahas

Terima kasih kepada dosen pembimbing dan pembahas yang telah membantu memberikan arahan, motivasi, dan ilmu yang berharga dalam proses penyusunan skripsi ini.

Almamater Tercinta Universitas Lampung

SANWACANA

Puji syukur kehadirat Allah SWT, atas segala rahmat dan karunianya sehingga penulis dapat menyelesaikan skripsi yang berjudul "Evaluasi Performa *Synthetic Minority Oversampling Technique* (SMOTE) untuk Mengatasi Klasifikasi Data Tidak Seimbang pada Metode *K-Nearest Neighbor* (KNN) dan *Support Vector Machine* (SVM)". Terselesaikannya skripsi ini tidak lepas dari dukungan, bimbingan, saran, serta doa dari berbagai pihak. Dengan segala kerendahan hati, penulis mengucapkan terima kasih kepada:

1. Ibu Dr. Khoirin Nisa, S.Si., M.Si., selaku Pembimbing I atas kesabaran dan seluruh kebaikan untuk bersedia memberikan kesediaan waktunya dalam memberikan arahan, bimbingan, bantuan dan dukungan yang sangat membangun sehingga penulis dipermudah dalam proses penyelesaian skripsi ini.
2. Ibu Dra. Dorrah Aziz, M.Si., selaku Pembimbing II yang telah memberikan waktunya untuk memberi bimbingan serta saran selama proses penyusunan skripsi ini.
3. Ibu Prof. Ir. Netti Herawati, M.Sc., Ph.D., selaku dosen Pembahas yang telah bersedia memberikan kritik dan saran serta evaluasi kepada penulis yang sangat membantu penulis dalam memperbaiki skripsi.
4. Bapak Dr. Ahmad Faisol, S.Si., M.Sc., selaku dosen Pembimbing Akademik yang telah membimbing penulis sampai akhir perkuliahan.
5. Bapak Dr. Aang Nuryaman, S.Si., M.Si., selaku Kepala Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung.
6. Bapak Dr. Eng. Heri Satria, S.Si., M.Si., selaku Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung.

7. Seluruh Dosen, Staf Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung yang telah memberikan ilmu, wawasan, dan pengetahuan yang berharga bagi penulis.
8. Papah Wahyu Supardan dan Mamah Ratu Entin Supartini selaku kedua orang tuaku tercinta yang telah memberikan cinta dan kasih sayang, kesabaran yang tulus ikhlas dalam membesarkanku yang tak ternilai harganya, merawat dan memberikan dukungan moral dan material serta doa yang tidak pernah putus selama penulis menempuh jenjang pendidikan S1 sehingga penulis dapat menyelesaikan jenjang pendidikan S1 ini. Semoga Allah senantiasa memuliakan Papah dan Mamah baik di dunia maupun di akhirat.
9. Adik-adikku tercinta, Daris Akmal Syafiq Al-Ghiffary, Keisha Azzira Triana Putri, dan Azzalea Thifalya Fami Anindhita yang selalu menghibur, memberikan semangat dan kasih sayang yang begitu besar terhadap penulis sehingga penulis menjadi termotivasi untuk menyelesaikan skripsi ini.
10. Sahabat-sahabatku tersayang, yaitu Melisa Saputri, Niken, Putri Dimar, Azzahra Zulfa Riswinda, Silvi Fitriani, dan Alfira Amalia Zuliyanti yang selalu menemani dikala suka maupun duka serta banyak memberikan dukungan dan motivasi dalam proses pengerjaan skripsi ini. Terima kasih sudah menjadi teman terbaik selama masa perkuliahan dan terima kasih sudah menutup cerita akhir perkuliahanku dengan sangat indah. Semoga kalian semua selalu diberikan kesuksesan dan kebahagiaan dalam hidup ini.
11. Teman seperbimbingan, Azizah Rahmahtia, Melisa Saputri, Ahmad Yusril Yusro, dan Nurjannah yang saling memberi dukungan dan saling menyemangati satu sama lain dalam proses penyelesaian skripsi ini.
12. Teman-teman Keluarga Besar Matematika 2019, terima kasih atas kebersamaannya selama 4 tahun masa perkuliahan.
13. Seluruh pihak yang tidak bisa penulis sebutkan satu-persatu secara detail yang telah membantu dan terlibat dalam menyelesaikan skripsi ini dengan baik.
14. Terakhir, terima kasih kepada Widya Amalia Putri Riswandha diri saya sendiri karena tidak pernah menyerah dan tetap berjuang dalam keadaan apapun. Mampu mengendalikan diri dari berbagai tekanan diluar keadaan dan tidak pernah memutuskan untuk menyerah sesulit apapun proses penyusunan

skripsi ini. Berada di titik ini bukanlah sesuatu yang mudah untuk dicapai dan ini merupakan suatu pencapaian yang patut dibanggakan untuk diri sendiri.

Semoga Allah SWT membalas segala kebaikan yang telah diberikan dengan cara sebaik-baiknya. Semoga skripsi ini dapat memberikan manfaat bagi para pembaca. Penulis menyadari bahwa terdapat banyak kekurangan dalam penulisan skripsi ini. Oleh karena itu, kritik dan saran sangat diharapkan guna menyempurnakan skripsi ini.

Bandar Lampung, 14 Agustus 2023
Penulis

Widya Amalia Putri Riswandha

DAFTAR ISI

Halaman

DAFTAR TABEL

Tabel Halaman

DAFTAR GAMBAR

Gambar Halaman

I. PENDAHULUAN

1.1 Latar Belakang dan Masalah

Salah satu metode yang populer dalam data *mining* dan *machine learning* adalah klasifikasi. Proses ini mencakup pencarian sebuah model yang menjelaskan dan membedakan konsep atau kelas data dengan tujuan memperkirakan kelas objek yang kelasnya tidak diketahui (Tan, dkk., 2006). Tujuan klasifikasi adalah untuk menemukan fungsi keputusan yang dapat memprediksi kelas data *testing* yang berasal dari fungsi distribusi yang sama dengan data untuk *training*. Terdapat beberapa metode dalam klasifikasi seperti *K-Nearest Neighbor* (KNN) dan *Support Vector Machine* (SVM). Prinsip kerja metode KNN adalah untuk menentukan nilai jarak pada data *testing* dengan data *training* menggunakan nilai terkecil dari nilai ketetanggaan terdekat (Satriya, dkk., 2018). Pada prinsipnya, metode SVM bekerja untuk menemukan sekumpulan pemisah optimal (*hyperplane*) dari data klasifikasi yang dilatih (Huang, dkk., 2014).

Pada umumnya, masalah yang sering terjadi pada klasifikasi yaitu ketidakseimbangan data. Ketidakseimbangan data terjadi ketika distribusi pada jumlah kelas tidak merata atau dapat diartikan bahwa terdapat perbedaan yang signifikan terhadap jumlah kelas (Naufal, dkk., 2015). Kelompok kelas yang memiliki jumlah data lebih banyak disebut kelompok mayoritas dan kelompok kelas yang memiliki jumlah data lebih sedikit disebut kelompok minoritas. Ketidakseimbangan data yang terjadi ini dapat mengakibatkan dampak yang kurang baik terhadap hasil klasifikasi dikarenakan

kelompok minoritas sering diklasifikasikan sebagai kelompok mayoritas (Siringoringo, 2018). Maka dari itu, penting untuk mengatasi masalah ketidakseimbangan data. Terdapat beberapa pendekatan yang digunakan untuk mengatasi permasalahan ketidakseimbangan data salah satunya yaitu dengan melakukan *resampling* (Chawla, dkk., 2002).

Teknik *resampling* secara umum terbagi menjadi dua yaitu *oversampling* dan *undersampling*. *Oversampling* adalah metode pembangkitan data kelas minoritas mendekati atau sama dengan kelas mayoritas (Chawla, 2009). Teknik *oversampling* memiliki beberapa metode, salah satunya yaitu *Synthetic Minority Oversampling Technique* (SMOTE). Metode SMOTE merupakan metode yang populer untuk menangani ketidakseimbangan kelas. Untuk menyeimbangkan data, metode ini bekerja dengan cara mensintesis sampel baru dari kelas minoritas (Siringoringo, 2018). SMOTE dapat menghasilkan nilai akurasi yang lebih tinggi dibandingkan dengan tanpa SMOTE. Proses evaluasi dilakukan untuk melihat pengaruh penggunaan SMOTE dalam mengatasi ketidakseimbangan data sebelum proses klasifikasi menggunakan algoritma KNN dan SVM.

Beberapa penelitian sebelumnya yang telah dilakukan, seperti Siringoringo (2018) menggunakan SMOTE dan KNN menghasilkan bahwa teknik SMOTE dapat menyelesaikan masalah ketidakseimbangan kelas pada *dataset Credit Card Fraud*. Penelitian yang dilakukan oleh Amelia, dkk. (2018) menggunakan metode SVM pada data tidak seimbang keberhasilan studi mahasiswa magister IPB. Penelitian ini memprediksi keberhasilan pemodelan SVM untuk studi mahasiswa dengan mempertimbangkan atribut dan latar belakang akademik mahasiswa. Dengan menggunakan SMOTE untuk menangani data tidak seimbang, pemodelan SVM berhasil meningkatkan kinerjanya dalam mengklasifikasikan mahasiswa yang tidak lulus. Penelitian lainnya yang dilakukan oleh Karlik, dkk. (2016) yang membahas bagaimana SMOTE digunakan untuk mengatasi masalah klasifikasi infertilitas dengan metode klasifikasi *Multi Level Perceptron* (MLP), KNN, *Naïve Bayes*, *Random Forest*, dan SVM, SMOTE berhasil meningkatkan performa klasifikasi.

Berdasarkan uraian di atas, maka pada penelitian ini akan mengkaji dan mengevaluasi performa *Synthetic Minority Oversampling Technique* (SMOTE) untuk mengatasi klasifikasi data tidak seimbang pada *K-Nearest Neighbor* (KNN) dan *Support Vector Machine* (SVM).

1.2 Tujuan Penelitian

Tujuan dari penelitian ini adalah untuk mengkaji dan mengevaluasi performa *Synthetic Minority Oversampling Technique* (SMOTE) untuk mengatasi klasifikasi data tidak seimbang pada *K-Nearest Neighbor* (KNN) dan *Support Vector Machine* (SVM).

1.3 Manfaat Penelitian

Manfaat dari penelitian ini adalah:

1. Menambah wawasan kepada penulis khususnya tentang mengatasi data kelas tidak seimbang menggunakan algoritma *Synthetic Minority Oversampling Technique*.
2. Dapat mengetahui performa *Synthetic Minority Oversampling Technique* pada *K-Nearest Neighbor* dan *Support Vector Machine* untuk mengatasi klasifikasi data tidak seimbang.
3. Penelitian ini diharapkan mampu menjadi bahan informasi untuk penelitian selanjutnya.

II. TINJAUAN PUSTAKA

2.1 Data *Mining*

Data *mining* merupakan proses mencari informasi berharga dalam sebuah data yang besar. Data *mining* adalah proses pencarian, analisis, dan penyaringan data yang besar untuk menemukan pola, tren, dan hubungan baru dalam data (Sumiran, 2018). Menurut definisi lain, data *mining* adalah proses berulang untuk menemukan pola atau model yang baru dan sempurna yang dapat dipahami dalam jumlah data yang sangat besar (Eska, 2018).

2.2 Tahapan Data *Mining*

Data *mining* memiliki tahapan-tahapan dalam pemrosesannya. Berikut merupakan tahapan-tahapan data *mining*:

1. Data *selection*, yaitu kumpulan *database* operasional yang dipilih atau diseleksi berdasarkan kebutuhan atau kepentingan sebelum melakukan proses data *mining*, kemudian disimpan dalam sebuah berkas atau tempat penyimpanan yang berbeda dengan *database* operasional sebelumnya agar mempermudah penggunaan data selanjutnya.
2. *Preprocessing*, yaitu membersihkan data dari isi yang tidak sempurna dari *database* seperti data yang hilang atau tidak valid, baik karena kesalahan pengetikan atau atribut yang tidak relevan agar tidak mengurangi nilai mutu

atau akurasi yang akan dihasilkan dari data tersebut. Dalam tahap ini performasi akan berpengaruh sebab data yang dihasilkan dapat berkurang jumlah dan kompleksitasnya.

3. *Transformation*, yaitu metode yang digunakan untuk mengubah bentuk format data sebelum memulai proses data *mining*. Kualitas proses ini ditentukan karena beberapa karateristik metode yang digunakan bergantung pada proses *transformation*. Pada proses *transformation* ini terdapat proses *scaling* data untuk membuat data numerik memiliki rentang yang sama.

   Terdapat dua cara yang digunakan dalam scaling data, yaitu:

   a. *Min Max Normalization* merupakan metode scaling data dengan melakukan transformasi linier terhadap data asli.

   $$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (2.1)$$

   dengan:

   x = nilai yang diamati

   x_{min} = nilai x minimum

   x_{max} = nilai x maximum

   b. *Z-Score Normalization (Standard Scaler)* merupakan teknik yang mana nilai pada atribut akan dinormalisasi berdasarkan mean dan standar deviasi.

   $$x_{standard} = \frac{x - \mu}{\sigma} \quad (2.2)$$

   dengan:

   x = nilai yang diamati

   μ = rata rata nilai (*mean*)

   σ = standar deviasi

4. Data *mining*, merupakan suatu proses yang dilakukan untuk menemukan model pola atau informasi menarik dari data yang telah di seleksi dan dilakukan dengan menggunakan teknik, metode, atau algoritma yang tepat sehingga sesuai dengan tujuan secara kesuruhan.
5. *Interpretation* (*Evaluation*), *interpretation* atau bisa disebut dengan *evaluation* adalah tahap menerjemahkan pola informasi yang dihasilkan dari pengolahan

data ke dalam format yang mudah dipahami oleh pihak yang berkepentingan. Selanjutnya, tahap ini memeriksa apakah pola informasi tersebut sesuai dengan fakta atau hipotesis yang telah ada.

2.3 *Machine Learning*

Pembelajaran mesin atau *machine learning* adalah bidang studi ilmiah yang menggunakan pola dan inferensi untuk melakukan tugas tertentu tanpa diberikan instruksi secara eksplisit. Algoritma dapat digunakan untuk menentukan kemampuan belajar menjadi dominan. Ini dapat dicapai melalui penggunaan kaidah, pendekatan statistik, atau pendekatan fisiologis. Untuk membuat prediksi atau keputusan, algoritma pembelajaran menggunakan data sampel atau yang lebih dikenal dengan *data training* untuk membangun model matematika (Bishop & Nasrabadi, 2006).

Algoritma *machine learning* terbagi menjadi tiga jenis, yaitu:

1. *Supervised Learning*

   Supervised learning merupakan *machine learning* yang dilakukan dengan adanya pelatihan dan pengujian sebagai pendekatan untuk menemukan suatu fungsi. Pengarahan atau pengawasan terhadap data latih yang bertujuan untuk menemukan fungsi atau hubungan, selanjutnya fungsi tersebut akan digunakan untuk data baru yang tidak berlabel (Kotu & Deshpande, 2015). *Supervised learning* diperlukan adanya data latih berlabel yang akan diterapkan fungsinya pada data tidak berlabel.
2. *Unsupervised Learning*

   Unsupervised learning merupakan algoritma yang tidak membutuhkan data berlabel. Data tersebut merupakan data yang tidak mempunyai atribut khusus didalamnya (Bramer, 2013).
3. *Reinforcement Learning*

   Reinforcement learning biasanya berada antara *supervised learning* dan *unsupervised learning*, metode ini bekerja dalam lingkungan yang dinamis di

mana konsepnya harus menyelesaikan tujuan tanpa diberitahu secara eksplisit oleh komputer jika tujuan tersebut telah tercapai (Das & Nene, 2017).

2.4 Data Tidak Seimbang

Data tidak seimbang terjadi ketika distribusi kelas data tidak merata atau ketika jumlah kelas data lebih banyak atau lebih sedikit daripada kelas yang lainnya (Ali, dkk., 2013). Kelompok kelas data yang memiliki jumlah yang lebih sedikit disebut kelompok minoritas dan kelompok kelas data yang memiliki jumlah yang lebih banyak disebut kelompok mayoritas.

Berikut macam-macam algoritma untuk mengatasi data tidak seimbang:

1. *Undersampling*, dilakukan dengan menyeimbangkan data pada kelas mayoritas dengan menjaga semua sampel pada kelas minoritas dan secara acak memilih jumlah sampel yang sama pada kelas mayoritas.
2. *Oversampling*, dilakukan dengan meningkatkan ukuran pada kelas minoritas.

2.5 *Oversampling*

Ketidakseimbangan data adalah salah satu permasalahan yang terjadi pada data *mining*. Salah satu penanganan ketidakseimbangan kelas adalah *resampling*. *Resampling* adalah salah satu teknik *preprocessing* dimana distribusi data diseimbangkan untuk mengurangi efek distribusi kelas tidak seimbang. Pendekatan *resampling* dibagi menjadi tiga, yaitu *oversampling*, *undersampling*, dan *hibrida* yang menggabungkan *oversampling* dan u*ndersampling*. *Oversampling* adalah metode pembangkitan data kelas minoritas agar mendekati atau sama dengan kelas mayoritas (Chawla, 2009). *Oversampling* bekerja untuk menyeimbangkan data dengan teknik acak tanpa menghapus pengamatan. Namun, *oversampling* bekerja dengan cara mereplikasi pengamatan pada data asli

yang dapat menyebabkan *overfitting* (Chawla, dkk., 2002). Salah satu teknik *oversampling* yaitu *Synthetic Minority Oversampling Technique* (SMOTE).

2.6 *Synthetic Minority Oversampling Technique* (SMOTE)

Synthetic Minority Oversampling Technique (SMOTE) merupakan salah satu teknik *oversampling* yang digunakan untuk mengatasi ketidakseimbangan data yang pertama kali diperkenalkan oleh Nithes V. Chawla pada tahun 2002. SMOTE merupakan metode penyeimbangan jumlah distribusi data sampel pada kelas minoritas dengan cara mensintesis data sampel tersebut hingga jumlah data sampel menjadi seimbang dengan jumlah sampel pada kelas mayoritas (Siringoringo, 2018).

Tahapan dalam melakukan penyeimbangan data dengan SMOTE dimulai dari menghitung jarak antar data pada kelas minoritas, kemudian menetukkan jumlah k terdekat, selanjutnya menghitung selisih antara data yang akan direplikasi dengan data dengan k tetangga terdekat dan yang terakhir adalah menciptakan data sintesis. Berikut langkah penerapan SMOTE (Chawla, 2002):

1. Menentukan data yang akan direplikasi (x_i) dari kelas minoritas yang dipilih secara random.
2. Menentukan nilai k (jumlah tetangga terdekat), kemudian menghitung jarak dari data x_i dengan data tetangga terdekat (x_{knn}) dalam kelas minoritas yang sama.
3. Setiap x_{knn} yang terpilih, hitung selisih antara x_i dan x_{knn}, lalu kalikan selisihnya dengan angka acak [0,1], dan tambahkan ke fitur yang diteliti.

$$x_{syn} = x_i + (x_{knn} - x_i) \times \delta \qquad (2.3)$$

dengan:

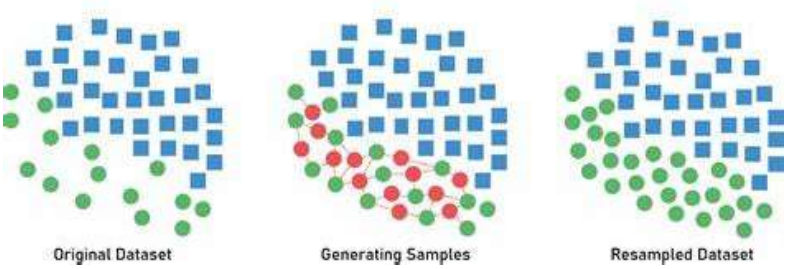
x_{syn} = data sintetis yang akan diciptakan

x_i = data yang akan direplikasi

x_{knn} = data yang memiliki jarak terdekat dari x_i

δ = nilai random dari [0,1]

4. Menggabungkan data asli dan data buatan.



Gambar 1. Ilustrasi penerapan SMOTE.

2.7 *K-Nearest Neighbor* (KNN)

Metode *K Nearest Neighbor* (KNN) dikembangkan oleh Evelyn Fix dan Joseph Hodges pada tahun 1995. Tujuan algoritma KNN adalah untuk mengklasifikasikan objek baru berdasarkan atribut dan data *training*. Algoritma ini mengklasifikasikan hasil *query instance* yang baru berdasarkan mayoritas kategori pada algoritma KNN. Kelas yang paling banyak muncul kemudian akan menjadi kelas hasil dari klasifikasi ini. Metode KNN menentukan nilai jarak data *testing* dengan data *training* dengan mengambil nilai terkecil dari nilai ketetanggaan terdekat. Setelah menghitung jarak antara data *testing* dan data *training*, hasil klasifikasi dihitung dari jarak terdekat. Data yang paling banyak masuk ke dalam anggota *k* sesuai dengan nilai *k* yang telah ditentukan sebelumnya dijadikan hasil klasifikasi (Prakoso, dkk., 2020). Pada KNN, terdapat tiga kunci elemen yakni banyaknya data *training*, nilai *k* yang digunakan, dan metode perhitungan jarak yang digunakan. Oleh karena itu perlu mencoba berbagai *k* agar didapatkan hasil klasifikasi yang terbaik (Wu, dkk., 2008).

Langkah-langkah dalam mengklasifikasi data menggunakan algoritma KNN yaitu:

a. Mendefinisikan nilai k.
b. Melakukan perhitungan nilai jarak atau *Euclidian* antara data *testing* dan data *training*.
c. Mengelompokkan data yang telah dilakukan perhitungan jarak berdasarkan nilai terkecil hingga nilai terbesar.
d. Memilih kelas yang paling banyak muncul dari sejumlah k yang dipilih untuk dijadikan sebagai hasil prediksi.

Berikut merupakan rumus untuk melakukan perhitungan jarak atau *Euclidian*

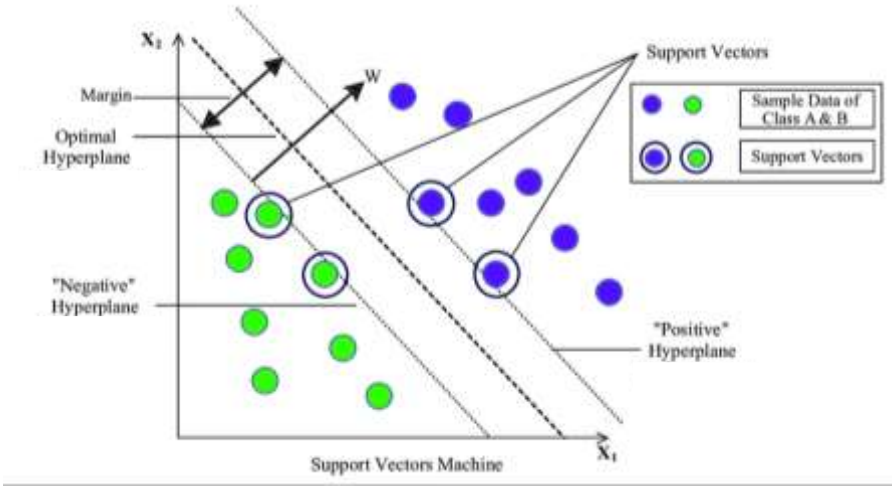
$$d(x_i, x_j) = \sqrt{\sum_{i,j=1}^{k}(x_i - x_j)^2} \quad (2.4)$$

dengan:

$d(x_i, x_j)$ = jarak *Euclidian*
k = banyaknya pengamatan
i, j = $1, 2, 3, \dots, k$

2.8 *Support Vector Machine* (SVM)

Support Vector Machine (SVM) merupakan salah satu *supervised learning* yang pertama kali diperkenalkan oleh Vapnik, dkk., pada tahun 1992. SVM adalah metode klasifikasi yang bekerja dengan cara mencari *hyperplane* terbaik untuk memisahkan dua kelas.



Gambar 2. *Support Vector Machine* menemukan *hyperplane* terbaik.

Hyperplane merupakan garis yang memisahkan data antar kelas. *Margin* diartikan sebagai jarak antara *hyperplane* dengan data terdekat pada maisng-masing kelas. Bidang pembatas pertama membatasi kelas pertama dan bidang pembatas kedua membatasi kelas kedua. Vektor-vektor pada bidang pembatas yang paling dekat dengan *hyperplane* terbaik disebut dengan *support vector*.

Untuk menemukan *hyperplane* sebagai pemisah terbaik antara dua kelas dilakukan pengukuran *margin* pada kedua kelas tersebut dan mencari titik maksimalnya. Misalkan data yang tersedia direpresentasikan dalam bentuk vektor:

$$D = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n); \boldsymbol{x_i} \in R, y_i \in \{-1,1\}$$

Data pada suatu *dataset* diberikan variabel x_i, sedangkan untuk kelas pada *dataset* diberikan variable y_i. Kelas pertama yang dipisah oleh *hyperplane* bernilai 1, sedangkan kelas lainnya bernilai -1. Maka persamaan yang didapatkan sebagai berikut:

$$\boldsymbol{x_i} . \boldsymbol{w^T} + b \geq +1 \text{ untuk } y_i = +1 \quad (2.5)$$

$$\boldsymbol{x_i} . \boldsymbol{w^T} + b \leq -1 \text{ untuk } y_i = -1 \quad (2.6)$$

dengan:

$\boldsymbol{x_i}$ = data ke i

y_i = label data kelas ke i

$\boldsymbol{w}$ = nilai bobot *support vector* yang tegak lurus dengan *hyperplane*

b = nilai bias

Memaksimalkan nilai jarak antara *hyperplane* dengan titik terdekatnya merupakan cara untuk menemukan nilai *margin* terbesar, yaitu $\frac{1}{\|\boldsymbol{w}\|}$. Hal ini dapat dirumuskan sebagai *Quadratic Programming* (QP) *problem*, yaitu dengan meminimalkan persamaan berikut:

$$\min_{w} \tau(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{w}\|^2 \tag{2.7}$$

dengan syarat

$$y_i(\boldsymbol{x_i} . \boldsymbol{w} + b) - 1 \geq 0, \forall i \tag{2.8}$$

Optimasi dapat dilakukan dengan menggunakan *lagrange multiplier* seperti berikut:

$$L = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{l} \alpha_i [y_i(\boldsymbol{x_i} . \boldsymbol{w^T} + b) - 1]$$

$$L = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{l} \alpha_i y_i(\boldsymbol{x_i} . \boldsymbol{w^T} + b) - \sum_{i=1}^{l} \alpha_i \tag{2.9}$$

α_i adalah *Lagrange Multipliers*, yang bernilai nol atau positif ($a_i \geq 0$). Nilai optimal dari persamaan di atas dapat dihitung dengan meminimalkan L terhadap w dan b, dan memaksimalkan L terhadap α_i.

$$\frac{\partial L}{\partial b} = 0$$

$$\sum_{i=1}^{l} \alpha_i y_i = 0 \tag{2.10}$$

$$\frac{\partial L}{\partial w} = 0$$

$$\boldsymbol{w} - \sum_{i=1}^{l} \alpha_i y_i \boldsymbol{x_i} = 0$$

$$\boldsymbol{w} = \sum_{i=1}^{l} \alpha_i y_i \boldsymbol{x_i} \tag{2.11}$$

Dengan memperhatikan sifat bahwa pada titik optimal gradient $L = 0$, persamaan di atas dapat dimodifikasi sebagai maksimalisasi masalah yang hanya mengandung α_i, sebagaimana persamaan berikut:

$$L = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{l} \alpha_i y_i (\boldsymbol{x_i} . \boldsymbol{w^T} + b) - \sum_{i=1}^{l} \alpha_i$$

$$L = \frac{1}{2}(\boldsymbol{w}.\boldsymbol{w^T}) - \left(\sum_{i=1}^{l} \alpha_i y_i \boldsymbol{x_i}\, \boldsymbol{w^T} + \sum_{i=1}^{l} \alpha_i y_i b - \sum_{i=1}^{l} \alpha_i \right)$$

$$L = \frac{1}{2}\left(\sum_{i=1}^{l} \alpha_i y_i \boldsymbol{x_i}. \sum_{j=1}^{l} \alpha_j y_j \boldsymbol{x_j} \right) - \left(\left(\sum_{i=1}^{l} \alpha_i y_i \boldsymbol{x_i} . \sum_{j=1}^{l} \alpha_j y_j \boldsymbol{x_j} \right) + 0 - \sum_{i=1}^{l} \alpha_i \right)$$

$$L = \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j \boldsymbol{x_i} \boldsymbol{x_j} - \left(\sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j \boldsymbol{x_i} \boldsymbol{x_j} - \sum_{i=1}^{l} \alpha_i \right)$$

$$L = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j \boldsymbol{x_i} \boldsymbol{x_j} \qquad (2.12)$$

dengan, $\alpha_i \geq 0, \sum_{i=1}^{l} \alpha_i y_i = 0$

Nilai α_i akan diperoleh dengan penyelesaian Persamaan (2.12) yang digunakan untuk mencari *primal variable* dengan rumus:

$$\boldsymbol{w_i} = \sum_{i=1}^{l} \alpha_i y_i K(\boldsymbol{x_i}, \boldsymbol{x_j})$$

$$b = -\frac{1}{2}(\boldsymbol{w^T x^+} + \boldsymbol{w^T x^-}) \qquad (2.13)$$

Setelah melakukan proses tersebut didapatkan nilai $\alpha_i > 0$ yang disebut dengan *support vector* dan sisanya memiliki nilai $\alpha_i = 0$. Fungsi keputusan yang dihasilkan hanya dipengaruh oleh nilai *support vector*.

2.8.1 Kernel *Trick*

Pada permasalahan data yang tidak dapat dipisahkan secara linier, maka digunakan *Non*-Linier *Support Vector Machine*. Metode *Non*-Linier SVM yang

dapat digunakan yaitu dengan pendekatan Kernel (Octaviani, dkk., 2014). Kernel *trick* yang dapat mengubah data *non*-linier menjadi linier (Hamel, 2009). Konsep kerja Kernel adalah dengan mentransformasi data ke dalam dimensi ruang fitur (*feature space*).

Beberapa fungsi Kernel yang umumnya digunakan dalam SVM adalah sebagai berikut:

a. Kernel Linier

$$K(\overrightarrow{x_i}, \overrightarrow{x_j}) = (\overrightarrow{x_i}, \overrightarrow{x_j}) \quad (2.14)$$

b. Kernel Polynomial

$$K(\overrightarrow{x_i}, \overrightarrow{x_j}) = (\overrightarrow{x_i}, \overrightarrow{x_j} + 1)^d \quad (2.15)$$

c. Kernel *Radial Basic Function* (RBF)

$$K(\overrightarrow{x_i}, \overrightarrow{x_j}) = \mathrm{e}^{-\left(\frac{\|\overrightarrow{x_i} - \overrightarrow{x_j}\|^2}{2\sigma^2}\right)}, \sigma > 0 \quad (2.16)$$

dimana $d, \gamma,$ dan σ merupakan parameter dari Kernel.

Penggunaan Kernel dapat dibedakan sesuai dengan data yang digunakan. Kernel linier digunakan pada saat data yang akan diklasifikasikan dapat dipisahkan oleh *hyperplane* berbentuk garis. Dalam artian lain, Kernel linier digunakan pada data berdimensi dua. Sebaliknya, Kernel *non*-linier digunakan pada data yang dipisahkan oleh *hyperplane* berbentuk bidang di ruang berdimensi tinggi (Puspitasari, dkk., 2018).

2.9 *Confusion Matrix*

Pada data *mining* terdapat berbagai cara untuk mengukur kinerja algoritma, salah satunya yaitu dengan *confusion matrix*. *Confusion matrix* menyatakan jumlah data *testing* yang benar dan jumlah data yang salah diklasifikasikan (Indriani, 2014).

Tabel 1. *Confusion Matrix*

Kelas Asli	Kelas Prediksi	
	Positif	Negatif
Positif	TP	FN
Negatif	FP	TN

dengan:

TP (*True Positive*) = jumlah data positif yang diprediksi positif
TN (*True Negative*) = jumlah data negatif yang diprediksi positif
FN (*False Negative*) = jumlah data negatif yang salah diprediksi sebagai positif
TN (*True Negative*) = jumlah data positif yang salah diprediksi sebagai negatif

Hasil nilai yang didapatkan dari *confusion matrix* dapat ditampilkan sebagai berikut:

1. *Accuracy*, merupakan nilai perbandingan antara data yang terklasifikasikan benar dengan keseluruhan data. Pengukuran ini digunakan untuk mengukur tingkat kebenaran klasifikasi (Hamel, 2009). Semakin tinggi nilai akurasi yang didapat maka semakin baik pula klasifikasi yang dihasilkan.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} x100\% \quad (2.17)$$

   Nilai akurasi klasifikasi tidak cukup sebagai ukuran kriteria standar, terutama pada kasus ketidakseimbangan antar kelas. Karena akan menghasilkan akurasi baik hanya untuk kelas mayoritas, sedangkan prediksi yang dihasilkan akan buruk untuk kelas minoritas.
2. *Precision*, yaitu untuk mengetahui seberapa sering model memberikan prediksi yang positif dan apakah prediksi itu benar dengan perumusan sebagai berikut:

$$Precision = \frac{TP}{TP+FP} \quad (2.18)$$

3. *Recall*, yaitu seberapa sering model memprediksi dengan benar pada data dengan klasifikasi aktual yang positif dengan perumusan sebagai berikut:

$$Recall = \frac{TP}{TP+FN} \quad (2.19)$$

4. *F1-score*, yaitu merupakan hubungan antara data berlabel positif dari hasil klasifikasi yang menunjukan keseimbangan antara *precision* dan *recall*

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (2.20)$$

III. METODOLOGI PENELITIAN

3.1 Waktu dan Tempat Penelitian

Penelitian ini dilakukan pada semester ganjil tahun 2022/2023 bertempat di Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung.

3.2 Data Penelitian

Data yang digunakan pada penelitian ini merupakan data sekunder mengenai penderita diabetes pada wanita yang berjumlah 768 data yang diambil dari situs kaggle.com sebagai berikut:

Tabel 2. Data Penelitian

No	Pregnancies	Glucose	BP	ST	Insulin	BMI	DPF	Age	Outcome
1	6	148	72	35	0	33.6	0.627	50	1
2	1	85	66	29	0	26.6	0.351	31	0
3	8	183	64	0	0	23.3	0.672	32	1
4	1	89	66	23	94	28.1	0.167	21	0
5	0	137	40	35	168	43.1	2.288	33	1
6	5	116	74	0	0	25.6	0.201	30	0
7	3	78	50	32	88	31	0.248	26	1
8	10	115	0	0	0	35.3	0.134	29	0
9	2	197	70	45	543	30.5	0.158	53	1

Tabel 2. (Lanjutan)

No	Pragnancies	Glucose	BP	ST	Insulin	BMI	DPF	Age	Outcome
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
768	1	93	70	31	0	30.4	0.315	23	0

Keterangan:

1. *Pregnancies* menjelaskan mengenai berapa kali responden mengalami kehamilan selama hidupnya
2. *Glucose* menjelaskan kadar glukosa dalam darah
3. *Blood Pressure* menjelaskan mengenai tekanan darah $\left(\frac{mm}{Hg}\right)$
4. *Skin Thickness* menjelaskan mengenai kadar lemak tubuh yang diukur pada tangan (mm)
5. *Insulin* menjelaskan mengenai tingkat insulin $\left(\frac{U}{ml}\right)$
6. *Body Mass Index* (BMI) menjelaskan mengenai indeks massa tubuh responden $\left(\frac{kg}{m^2}\right)$
7. *Diabetes Pedigree Function* menjelaskan indikator riwayat diabetes dalam keluarga
8. *Age* menjelaskan umur responden
9. *Outcome* menjelaskan apakah responden tersebut merupakan penderita diabetes yang diberi label 1 dan responden yang bukan merupakan penderita diabetes diberi label 0

3.3 Metode Penelitian

Langkah-langkah yang dilakukan pada penelitian ini sebagai berikut:

1. Melakukan input data.
2. Melakukan analisis deskriptif yang bertujuan untuk melihat ringkasan data dari variabel penelitian.
3. Melakukan *preprocessing* data.
    a. Melakukan *cleaning* data untuk memeriksa apakah terdapat data hilang atau data duplikat.
    b. Melakukan standarisasi data.
    c. Membagi data dengan rasio 70% data *training* 30% data *testing*, 80% data *training* 20% data *testing*, 90% data *training* 10% data *testing*.
4. Melakukan proses klasifikasi KNN dan SVM dengan data yang tanpa diberikan perlakuan SMOTE.
    a. Proses Klasifikasi KNN
        - Mendefinisikan nilai k.
        - Melakukan perhitungan nilai jarak atau *Euclidian* antara data *testing* dan data *training*.
        - Mengelompokkan data yang telah dilakukan perhitungan jarak berdasarkan nilai terkecil hingga nilai terbesar.
        - Memilih kelas yang paling banyak muncul dari sejumlah k yang dipilih untuk dijadikan sebagai hasil prediksi.
    b. Proses Klasifikasi SVM
        - Menentukan fungsi Kernel yang akan digunakan.
        - Menentukan parameter pada fungsi Kernel yang digunakan.
        - Membangun model SVM menggunakan fungsi Kernel.
        - Membentuk *confusion matrix* dan menghitung performa klasifikasi berdasarkan ukuran akurasi, *precision*, *recall*, dan *F1-score*.

5. Melakukan proses klasifikasi KNN dan SVM dengan data yang telah diberikan perlakuan SMOTE.
    a. Menentukan data yang akan direplikasi (x_i) dari kelas minoritas yang dipilih secara random.
    b. Menentukan nilai k (jumlah tetangga terdekat), kemudian menghitung jarak dari data x_i dengan data tetangga terdekat (x_{knn}) dalam kelas minoritas yang sama.
    c. Setiap x_{knn} yang terpilih, hitung selisih antara x_i dan x_{knn}, lalu kalikan selisihnya dengan angka acak [0,1], dan tambahkan ke fitur yang diteliti.
    d. Menggabungkan data asli dan data buatan.
6. Melakukan evaluasi hasil klasifikasi KNN dan SVM dengan SMOTE dan tanpa SMOTE.

V. KESIMPULAN

Setelah melakukan proses *machine learning* dengan menggunakan *Synthetic Minority Oversampling Technique* (SMOTE) yang digunakan pada metode *K Nearest Neighbor* (KNN) dan *Support Vector Machine* (SVM) untuk menyeimbangkan data pada klasifikasi penderita diabetes, dapat diambil kesimpulan bahwa SMOTE efektif untuk meningkatkan performa klasifikasi penderita diabetes. Hal ini dapat dilihat dari adanya peningkatan nilai akurasi pada data yang telah diberikan perlakuan SMOTE. Pada hasil klasifikasi KNN dengan skema data 90% data *training* dan 10% data *testing* diperoleh nilai akurasi sebesar 75% dan setelah diberikan perlakuan SMOTE nilai akurasi meningkat menjadi 81.58%. Pada hasil klasifikasi SVM dengan skema data 90% data *training* dan 10% data *testing* dengan menggunakan Kernel RBF diperoleh nilai akurasi sebesar 83.12% dan setelah diberikan perlakuan SMOTE nilai akurasi meningkat menjadi 85.53%. Oleh karena itu, SMOTE dinilai efektif dalam meningkatkan performa klasifikasi penderita diabetes pada metode KNN dan SVM.

DAFTAR PUSTAKA

Ali, A., Shamsuddin, S.M., & Ralescu, A.L. 2013. Classification with Class Imbalance Problem. *International Journal Advance Soft Computer Application*. **5**(3): 2-7.

Amelia, O.D., Soleh, A.M., & Rahardiantoro, S. 2018. Pemodelan Support Vector Machine Data Tidak Seimbang Keberhasilan Studi Mahasiswa Magister IPB. *Institute of Electrical and Electronics Engineers*. **2**(1): 33-40.

Bishop, C.M. & Nasrabadi, N.M. 2006. *Pattern Recognition and Machine Learning*. Springer, New York.

Bramer, M. 2013. *Principles of Data Mining*. Springer, London.

Chawla, N.V., Bowyer, K.W., Hall, L.O., & Kegelmeyer, W.P. 2002. SMOTE: Synthetic Minority OverSampling Technique. *Journal of Artificial Intelligence Research*. **1**(6): 321-357.

Chawla, N.V. 2009. *Data Mining for Imbalanced Datasets: an Overview Data Mining and Knowledge Discovery Handbook*. Springer, Berlin.

Das, S. & Nene, M.J. 2017. A Survey on Types of Machine Learning Techniques in Intrusion Prevention Systems, hlm. 2296-2299. International Conference on Wireless Communications.

Eska, J. 2018. Penerapan Data Mining Untuk Prediksi Penjualan Wallpaper Menggunakan Algoritma C45. *Jurnal Teknologi dan Sistem Informasi*. **2**(2): 2-3.

Gorunescu, F. 2011. *Data Mining : Concept, Model and Techniques*. Springer. Berlin.

Hamel, L. 2009. *Model Assessment with ROC Curves*: *Encyclopedia of Data Warehousing and Mining*. 2^{nd} Edition. IGI Global, Pennsylvania.

Huang, M.L., Hung, Y.H., Lee, W.M., Li, R.K., & Jiang, B.R. 2014. SVM-RFE Based Feature Selection and Taguchi Parameters Optimization for Multiclass SVM Classifier. *The Scientific World Journal*. **2**(4): 1-10.

Indriani, A. 2014. Klasifikasi Data Forum dengan menggunakan Metode Naïve Bayes Classifier, hlm. 5-9. Prosiding Seminar Nasional Aplikasi Teknologi Informasi.

Karlik, B., Yibre, A.M., & Koçer, B. 2016. Comprising Feature Selection and Classifier Methods with SMOTE for Prediction of Male Infertility. *International Journal Fuzzy Syst. Adv. Appl.* **3**(1): 1-6.

Kotu, V. & Deshpande, B. 2014. *Predictive Analytics and Data Mining: Concepts and Practice with Rapidminer*. Morgan Kaufmann, Burlington.

Naufal, A.R., Wahono, R.S., & Syukur, A. 2015. Penerapan Bootstrapping untuk Ketidakseimbangan Kelas dan Weighted Information Gain untuk Feature Selection pada Algoritma Support Vector Machine untuk Prediksi Loyalitas Pelanggan. *Journal of Intelligent Systems*. **1**(2): 98-108.

Octaviani, P.A., Wulandari, Y., & Ispriyanti, D. 2014. Penerapan Metode Klasifikasi Support Vector Machine (SVM) pada Data Akreditasi Sekolah Dasar (SD) di Kabupaten Magelang. *Jurnal Gaussian*. **3**(8): 811–820.

Prakoso, R.D.Y., Wiriaatmadja, B.S., & Wibowo, F.W. 2020. Sistem Klasifikasi pada Penyakit Parkinson dengan Menggunakan Metode K-Nearest Neighbor, hlm. 63-68. Prosiding Seminar Nasional Teknologi Komputer dan Sains.

Puspitasari, A.M., Ratnawati, D.E., & Widodo, A.W. 2018. Klasifikasi penyakit gigi dan mulut menggunakan metode Support Vector Machine. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*. **2**(2): 802-810.

Satriya, R.H.D., Santoso, E., & Sutrisno. 2017. Implementasi Metode Ensemble K-Nearest Neighbor untuk Prediksi Nilai Tukar Rupiah Terhadap Dollar Amerika. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*. **2**(4): 1718-1725.

Siringoringo, R. 2018. Klasifikasi Data Tidak Seimbang Menggunakan Algoritma SMOTE dan K-Nearest Neighbor. *Journal Information System Development.* **3**(1): 44–49.

Sumiran, K. 2018. An Overview of Data Mining Techniques and Their Application in Industrial Engineering. *Asian Journal of Applied Science and Technology*. **2**(2): 947-953.

Tan, P., Steinbach, M., & Kumar, V. 2006. *Introduction to Data Mining*. KMedia, Yogyakarta.

Wu, H., Yang, S., Huang, Z., He, J., & Wang, X. 2018. Type 2 Diabetes Mellitus Prediction Model Based on Data Mining. *Informatics in Medicine Unlocked*. **2**(10): 100-107.